NAME

upmendex – Multilingual index processor

SYNOPSIS

upmendex [-ilqrcgf] [**-s** *sty*] [**-d** *dic*] [**-o** *ind*] [**-t** *log*] [**-p** *no*] [**--**] [*idx0 idx1 idx2 ...*]
upmendex --help

DESCRIPTION

The program *upmendex* is a general purpose multilingual hierarchical index generator working with upLaTeX, XeLaTeX and LuaLaTeX; it accepts one or more input files (*.idx*; often produced by a text formatter such as LaTeX families), sorts the entries, and produces an output file which can be formatted. It supports Latin (including non-English), Greek, Cyrillic, Korean Hangul and Han (Hanzi ideographs) scripts, as well as Japanese Kana. It is almost compatible with *makeindex* and *mendex*, and additional feature for handling readings of kanji words is also available.

The formats of the input and output files are specified in a style file. The readings of kanji words can be specified in a dictionary file.

The index can have up to five levels (0, 1, 2, 3, and 4) of subitem nesting.

OPTIONS

-i Take input from stdin, even when index files are specified.

-l Set ´sort by character order´. By default, ´sort by word order´ is used. Details are described below.

-q Quiet mode; send no message to *stderr*, except error messages and warnings.

-r Disable implicit page range formation. By default, three or more successive pages are automatically abbreviated as a range (e.g. 1–5).

-c Compress sequence of intermediate blanks (space(s) and/or tab(s)) into a space and ignore leading and trailing blank(s). By default, blanks in the index key are retained.

-g Make Japanese index head *A*-line (*A, Ka, Sa, ...*; 10 characters) of the *gojuon* table (Japanese syllabary). By default, all 48 characters in the *gojuon* table are used.

-f Force to output characters even if the scripts are not supported by *upmendex*.

-s *sty* Employ *sty* as the style file.

-d *dic* Employ *dic* as the dictionary file. The dictionary file is composed of lists of <*index_word reading*>.

-o *ind* Employ *ind* as the output index file. By default, the file name is created by appending the extension *ind* to the base name of the first input file.

-t *log* Employ *log* as the transcript file. By default, the file name is created by appending the extension *ilg* to the base name of the first input file.

-p *no* Set the starting page number of the output index list to be *no*. The argument *no* may be numerical or one of the following: *any* (the next page to the end of contents), *odd* (the next odd page to the end of contents), *even* (the next even page to the end of contents).

--help Show summary of options.

-- Arguments after **--** are not taken as options. This is useful when the input file name starts with '-'.

STYLE FILE

The style file informs *upmendex* about the format of the *idx* input files and the intended format of the final output file. The format is upper compatible with the one for *makeindex* and *mendex*. The style file contains a list of <*specifier attribute*> pairs. There are two types of specifiers: input and output. Pairs do not have to appear in any particular order. A line begun by ´%´ is a comment.

Input file style parameter

keyword <string>
"\\indexentry"
Command with an argument of index entry which is going to be processed.

arg_open <char>
´{´
Opening delimiter which shows the beginning of index entry.

arg_close <char>
´}´
Closing delimiter which shows the end of index entry.

range_open <char>
´(´
Opening delimiter which shows the beginning of page range.

range_close <char>
´)´
Closing delimiter which shows the end of page range.

level <char>
´!´
Delimiter which shows lower level.

actual <char>
´@´
Symbol which shows the next sequence is to appear as index strings in the output file.

encap <char>
´|´
Symbol which shows the next sequence is to be used as command name attached to the page number.

page_compositor <string>
"-"
Separator between page levels for a style with multi-levels of page numbers.

page_precedence <string>
"rnaRA"
Priority of expression for page number. ´R´ and ´r´ correspond to Roman. ´n´ corresponds to arabic numeral. ´A´ and ´a´ correspond to Latin alphabet.

quote <char>
´"´
Escape character for *upmendex* parameters.

escape <char>
´\\´
Escape character for general scripts.

Output file style parameter

preamble <string>
"\\begin{theindex}\n"
Preamble of output file.

postamble <string>
"\n\n\\end{theindex}\n"
Postamble of output file.

setpage_prefix <string>
"\n \\setcounter{page}{"
Prefix of page number if start page is designated.

setpage_suffix <string>
"}\n"
Suffix of page number if start page is designated.

group_skip <string>
"\n\n \\indexspace\n"
Strings to insert vertical space before new section of index.

lethead_prefix <string>
""
Prefix of heading for newly appeared heading letter.

heading_prefix <string>
""
Same as **lethead_prefix**. (compatible with makeindex)

lethead_suffix <string> ""
Suffix of heading for newly appeared heading letter.

heading_suffix <string> ""
Same as **lethead_suffix**. (compatible with makeindex)

lethead_flag <number> 0
Flag to control output of heading letters in Latin, Greek and Cyrillic scripts. ´0´, ´1´, ´-1´ and ´2´ respectively denotes no output, uppercase, lowercase and titlecase.

heading_flag <number> 0
Same as **lethead_flag**. (Note: makeindex uses a different name **headings_flag**)

headings_flag <number> 0
Same as **lethead_flag**. (compatible with makeindex)

kana_head <string> ""
Heading characters of Kana specified by a string. By default, it is controlled by **letter_head** and command line option **-g**. (Extended by upmendex)

hangul_head <string> "111417191A1B1E1G1H1J1K1L1M1N"
Heading characters of Hangul specified by a string. (Extended by upmendex)

tumunja <string> "111417191A1B1E1G1H1J1K1L1M1N"
Heading characters of Hangul specified by a string. (Deprecated, Extended by upmendex)

hanzi_head <string> ""
Heading strings of hanzi (Kanji, Hanja) specified by a string, which is concatenated of items with a separator ´;´. (Extended by upmendex)

devanagari_head <string> "(Devanagari script)"
Heading characters of Devanagari specified by a string. (Experimental, Extended by upmendex)

thai_head <string> "(Thai script)"
Heading characters of Thai script specified by a string. (Experimental, Extended by upmendex)

item_0 <string> "\n \\item "

item_1 <string> "\n \\subitem "

item_2 <string> "\n \\subsubitem "

item_3 <string> "\n \\subsubsubitem "

item_4 <string> "\n \\subsubsubsubitem "
Command sequence inserted between two primary, two secondary, etc. level entries.

item_01 <string> "\n \\subitem "

item_12 <string> "\n \\subsubitem "

item_23 <string> "\n \\subsubsubitem "

item_34 <string> "\n \\subsubsubsubitem "
Command sequence inserted between primaly and secondary, secondary and tertiary, etc. level entries.

item_x1 <string> "\n \\subitem "

item_x2 <string> "\n \\subsubitem "

item_x3 <string> "\n \\subsubsubitem "

item_x4 <string> "\n \\subsubsubsubitem "
Command sequence inserted between primary and secondary, secondary and tertiary, etc. level entries when the higher level entry does not have page number.

delim_0 <string> ", "

delim_1 <string> ", "

delim_2 <string> ", "

delim_3 <string> ", "

delim_4 <string> ", "
Delimiter string between primary, secondary, etc. level entry and first page number.

delim_n <string> ", "
Delimiter string between page numbers commonly used for any entry level.

delim_r <string> "--"
Delimiter string between pages to show page range.

delim_t <string> ""
Delimiter string output at the end of page number list.

suffix_2p <string> ""
String to be inserted in place of **delim_n** and the next page number when the two pages are contiguous.
It works only when the parameter is defined.

suffix_3p <string> ""
String to be inserted in place of **delim_r** and the third page number when the three pages are contiguous. The parameter is prior to **suffix_mp**.
It works only when the parameter is defined.

suffix_mp <string> ""
String to be inserted in place of **delim_r** and the last page number when the three or more pages are contiguous.
It works only when the parameter is defined.

encap_prefix <string> "\\"
Prefix for an encapsulating command when the encapsulating command is added to the page number.

encap_infix <string> "{"
Prefix just before the page number when the encapsulating command is added to the page number.

encap_suffix <string> "}".
Suffix after the page number when the encapsulating command is added to the page number.

line_max <number> 72
Maximum number of one line. If exceed the number, lines are folded.

indent_space <string> ""
Space for indent which inserted to top of folded line.

indent_length <number> 16
Length of space for indent which inserted to top of folded line.

symhead_positive <string> "Symbols"
Strings to output as heading letter for symbols when lethead_flag or heading_flag or headings_flag is positive number.

symhead_negative <string> "symbols"
Strings to output as heading letter for symbols when lethead_flag or heading_flag or headings_flag is negative number.

symbol <string> ""
Strings to output as heading letter for symbols when symbol_flag is non zero.
If specified, the option is prior to symhead_positive and symhead_negative. (Extended by (up)mendex)

numhead_positive <string> "Numbers"
Strings to output as heading letter for numbers when lethead_flag or heading_flag or headings_flag is positive number and symbol_flag is 2.

numhead_negative <string> "numbers"
Strings to output as heading letter for numbers when lethead_flag or heading_flag or headings_flag is negative number and symbol_flag is 2.

symbol_flag <number> 1
Flag to output of symbol. If ´0´, do not output headings for symbols and numbers. If ´1´, output symbols and numbers as a group of symbols. If ´2´, output symbols and numbers separately. (Extended by (up)mendex)

letter_head <number> 1
Flag of heading letter for Japanese Kana. If ´1´ and ´2´, Katakana and Hiragana is used, respectively. (Extended by (up)mendex)

priority <number> 0
Flag of sorting method for index words composed of Japanese and non-Japanese (ex. Latin scripts). If non zero, one space (U+0020) is inserted between Japanese sequence and non-Japanese sequence in sorting procedure. (Extended by (up)mendex)

character_order <string> "SNLGCJKHDTah"
Order of scripts and symbols. ´S´, ´N´, ´L´, ´G´, ´C´, ´J´, ´K´, ´H´, ´D´, ´T´, ´a´ and ´h´ respectively denotes symbol, number, Latin, Greek, Cyrillic, Japanese Kana, Korean Hangul, Hanzi, Devanagari, Thai, Arabic and Hebrew script. ´@´ denotes scripts which are not explicitly designated and the order are configured by icu_rules or icu_locale. Please make sure that ´S´ and ´N´ are next to each other if symbol_flag=1, since numbers are classified as a part of symbol. (Extended by upmendex)

script_preamble <string 1> <string 2>
""
Preamble of script block in output file, specified by string 2. One of script names must be specified in the string 1: ´latin´, ´cyrillic´, ´greek´, ´kana´, ´hangul´, ´hanzi´, ´devanagari´, ´thai´, ´arabic´, or ´hebrew´. (Extended by upmendex)

script_postamble <string 1> <string 2>
""
Postamble of script block in output file, specified by string 2. One of script names must be specified in the string 1: ´latin´, ´cyrillic´, ´greek´,

´kana´, ´hangul´, ´hanzi´, ´devanagari´, ´thai´, ´arabic´, or ´hebrew´. (Extended by upmendex)

icu_locale <string> ""

Locale in ICU collator. By default, "root sort order" is set. (Extended by upmendex)

icu_rules <string> ""

Customized collation rules in ICU collator. Unicode characters in UTF-8 encoding and following escape sequences are accepted: **\Uhhhhhhhh** (8-digit hexadecimal [0-9A-Fa-f]), **\uhhhh** (4-digit hexadecimal), **\xhh** (2-digit hexadecimal), **\x{h...}** (1..8-digit hexadecimal), and **\ooo** (3-digit octal [0-7]). If icu_rules and icu_locale are simultaneously specified, collation rules specified by icu_rules are added on collation rules specified by icu_locale. By default, locale is used. (Extended by upmendex)

Ref. <https://unicode-org.github.io/icu/userguide/collation/customization/>, <http://www.unicode.org/reports/tr35/tr35-collation.html#Rules>

icu_attributes <string> ""

Attributes in ICU collator. Followings are available: "alternate:shifted", "alternate:non-ignorable", "strength:primary", "strength:secondary", "strength:tertiary", "strength:quaternary", "strength:identical", "french-collation:on", "french-collation:off", "case-first:off", "case-first:upper-first", "case-first:lower-first", "case-level:on", "case-level:off", "normalization-mode:on", "normalization-mode:off", "numeric-ordering:on", "numeric-ordering:off" (Extended by upmendex)

Ref. <https://unicode-org.github.io/icu/userguide/collation/customization/#default-options>, <http://www.unicode.org/reports/tr35/tr35-collation.html#Setting_Options>

ABOUT JAPANESE PROCESSING

upmendex has an additional feature to simplify the procedure of handling Japanese indexes, compared to *makeindex*. Users can save the effort of manually specifying a reading for every kanji word.

Japanese kanji words are usually sorted by the syllables of their readings (*´Yomi´*), which can be represented by kana (Hiragana, Katakana) scripts. *upmendex* accepts index words specified in kana expression directly on an input file, and also accepts conversion from index words in Kanji or symbols to phonogram scripts by referring to Japanese dictionaries.

Examples of internal simplification of syllables are shown below.

かぶしきがいしゃ	かふしきかいしや
マッキントッシュ	まつきんとつしゆ
ワープロ	わあふろ

The dictionary file consists of list with <´index_word´ ´reading´>. The index word can be written in any scripts (kanji, kana, etc), and the reading can be in any phonograms such as Hiragana or Katakana scripts. The delimiter between the index word and its reading is one or more tab(s) or space(s).
An example of a Japanese dictionary is shown below.

漢字	かんじ
読み	よみ
環境	かんきょう
$	ドル

Here, each index word is allowed to have only one Yomi. Though some kanji words (ex.「表」) may have more than one Yomi´s (ex.「ひょう」and「おもて」), only one of them can be registered in the dictionary. When some different Yomi´s are needed, they should be specified explicitly in kana expression (ex. \index{ひょう@表} or \index{おもて@表}) on the input file.
Moreover, a dictionary file is automatically referred by setting the file name at an environment variable

INDEXDEFAULTDICTIONARY. The dictionary set by the environment variable can be used together with file(s) specified by *-d* option.

ABOUT SORTING PROCEDURE

upmendex sorts indexes as is (´sort by word order´) by default. Setting *-l* option, spaces between words in an index are truncated prior to sorting procedure (´sort by character order´).

Even when sort by character order, the index at output remains the original sequence without the truncation. Follows show an example.

sort by word order	*sort by character order*
X Window	Xlib
Xlib	XView
XView	X Window

In addition, two sorting methods can be applied for indexes which contains both Japanese kana and other scripts (e.g. Latin script). By setting *priority* 0 (default) and 1 at a style file, a space between Japanese Kana and other scripts is inserted and not inserted respectively, prior to the sorting procedure.
Follows show an example.

priority=0	*priority=1*
index sort	indファイル
indファイル	index sort

ENVIRONMENT VARIABLES

upmendex refers environment variables as follows.

INDEXSTYLE
: Directory where index style files exist.

INDEXDEFAULTSTYLE
: Index style file to be referred to as default.

INDEXDICTIONARY
: Directory where dictionary files exist.

INDEXDEFAULTDICTIONARY
: Dictionary file which is automatically read.

DETAIL

Detailed specification is compatible with *makeindex*.

KNOWN ISSUES

When plural page number expression is used, *.idx* files should be specified along with the order of page numbers. Otherwise, wrong page numbers might be output.

SEE ALSO

tex(1), latex(1), makeindex(1), mendex(1).
International Components for Unicode (ICU): <http://icu.unicode.org/>, <https://unicode-org.github.io/icu/>

AUTHOR

This manual page was written by Takuji Tanaka based on the mendex manual page written by Japanese TEX Development Community.